Asian Journal of Information Technology 5 (11): 1172-1176, 2006

A Segment Based Approach of Hidden Markov Models for Speech Recognition

Rafik Djemili, Mouldi Bedda and Hocine Bourouba
Département d'Electronique, Faculté des Sciences de l'Ingénieur,
Université Badji Mokhtar de Annaba, BP 12 Sidi Amar, Annaba 23000, Algérie

Abstract: In this study propose a new approach in using Hidden Markov Models (HMMs) for speech recognition. Although HMMs are the state-of-the art speech recognition systems, they suffer from some inherent limitations. One of these limitations is the independence assumption in the HMMs formalism. In the approach described in this study, we use in the vector quantization process, grouped vectors of different length to explicitly model the natural correlation between adjacent frames, instead of using a single vector in the standard method. The system is tested on an Arabic isolated digits (0-9) recognition task, our method achieves a 21% reduction in word error rate evaluation compared with the standard approach.

Key words: Speech recognition-hidden markov models-vector quantization-segment models-grouped vectors

INTRODUCTION

To date, the most successful speech recognition systems have been based on Hidden Markov Models (HMMs)[1-3] and the use of HMMs for acoustic modelling dominates the continuous speech recognition field. Although HMMs will continue to play a role in most speech recognition systems for a long time to come, they suffer from major limitations which handicapped them to reach human performance in tasks related to speech recognition[4]. One of these limitations is the independence assumption which says that there is no correlation between adjacent input frames, so HMMs examine only one frame of speech at a time[5]. To overcome this weakness in the HMMs formalism, many researchers developed ideas such as augmenting the observation space with feature derivatives[6], a viable solution but doesn't resolve the problem completely, or proposed to explicitly model correlation, including conditionally Gaussian HMMs[7] and segmental HMMs[8,9]. However, these approaches known in the literature as Segment Models (SMs) while very powerful, tend to make changes to the standard HMMs training and recognition algorithms and with a higher computational cost due to the expanded state space[10].

In this study, we propose a new approach tackling the correlation between adjacent input frames without changing any algorithm of the standard HMMs. The idea behind our approach is to group frames (vectors) given by the feature extraction and treat them as a single observation vector in the Vector Quantization (VQ) process.

HIDDEN MARKOV MODELS

Acoustic modeling: The goal of acoustic modeling is to derive some convenient representation of speech signals before their use in a speech recognition system. Hence each speech signal in the current study, is sampled at 22050 Hz decimated at a rate of 11025 Hz, passes through a high frequency preemphasis filter with a transfer function $H(z) = 1 - az^{-1}$. The preemphasized data is blocked into overlapping frames. Each frame is 23.2 ms duration, with 11.6 ms spacing. Spectral analysis is performed to get twelve Mel Frequency Cepstral Coefficients (MFCC)[11] and the log of the energy calculated in the temporal domain. The first twelve MFCC are obtained from the energies of F bank filters directly using the DCT transform:

$$c_k = \sum_{i=1}^{F} \log E_i \cos\left[\frac{\pi k}{F}(i - \frac{1}{2})\right] 1 \le k \le d \quad (1)$$

The overall system is depicted in Fig. 1:

Presentation of hidden markov models: In this subsection, we remind the basic definition of an HMM, we formalize the assumptions that are made and describe the basic elements of algorithms for HMMs, we use the notation as in[2]. A hidden Markov model can be defined as a doubly embedded stochastic process with an underlying stochastic process that is not observable (it is hidden) but can only be observed through another set of stochastic processes that produce the sequence of observations.

Corresponding Author: Rafik Djemili, Département d'Electronique, Faculté des Sciences de l'Ingénieur, Université Badji Mokhtar de Annaba, BP 12 Sidi Amar, Annaba 23000, Algérie

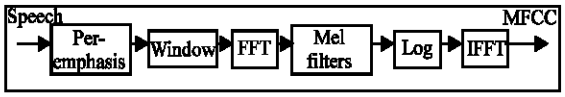

Fig. 1: Analysis of speech frames

A Markov model of order k is a probability distribution over a sequence of variables $q_1^t = \{q_1, q_2, ..., q_t\}$ with the following conditional independence property:

$$p\left(q_t \middle| q_1^{t-1}\right) = p\left(q_t \middle| q_{t-k}^{t-1}\right) \tag{2}$$

Since q_{t-k}^{t-1} summarizes all the relevant past information, q_t is generally called a state variable. Because of the above conditional independence property, the joint distribution of a whole sequence can be decomposed into the product:

$$p\left(q_1^T\right) = p\left(q_1^K\right) \prod_{t=k+1}^{T} p\left(q_t \middle| q_{t-k}^{t-K}\right) \tag{3}$$

The special case of a Markov model of order 1 is the one used in our study. In this case, the distribution is even simpler:

$$p\left(q_1^T\right) = p\left(q_1\right) \prod_{t=2}^{T} p\left(q_t \middle| q_{t-K}\right) \tag{4}$$

and it is completely specified by the so-called initial state probabilities $p(q_1)$ and transition probabilities $p(q_t|q_{t-1})$ An HMM is characterized by the following five elements:

- N: The number of states in the model.
- M: The number of distinct observation symbols per state, we denote an observation sequence by:

$$o = \{o_1, o_2, ..., o_T\}$$

- The state transition probability distribution $A = \{a_{ij}\}$ where

$$a_{ij} = \text{prob}[q_{t+1} = S_j / q_t = S_i] \; 1 \le i, j \le N \tag{5}$$

i.e.: The probability of being in state S_j at time t+1 and in state S_i at time t.

- The observation symbol probability distribution in state j, $B = \{b_j(k)\}$ where

$$b_j(k) = \text{prob}[v_k \text{ at } t / q_t = S_j] \quad 1 \le k \le M \tag{6}$$

i.e.: The probability of observing the symbol v_k at time t in the state S_j.

- The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = \text{prob}(q_1 = S_i) \quad 1 \le i \le N \tag{7}$$

i.e.: The probability of being in state S_j at time t=1.

For convenience, we use the compact notation $\lambda = (A,B,\pi)$ to indicate the complete parameter set of the model.

Viterbi algorithm: To find the single best state sequence $Q = \{q_1, q_2, ..., q_T\}$ for a given observation sequence $o = \{o_1, o_2, ..., o_T\}$, we use a formal technique based on dynamic programming methods and called the Viterbi algorithm[12]. We first define the quantity:

$$\delta_t(i) = \max_{q_1 \ldots q_T} p\left[q_1 \ldots q_t = i, o_1 \ldots o_t \middle| \lambda\right] \tag{8}$$

i.e.: $\delta_t(i)$ is the best score (highest probability along a single path, at time t, which accounts for the first t observations and ends in state S_i. By induction we have:

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] b_j(o_{t+1}) \tag{9}$$

To actually retrieve the state sequence, we need to keep track of the argument which maximized (9) for each t and j. We do this via the array $\psi_t(j)$. The complete procedure for finding the best sequence can now be stated as follows:

Initialization:

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(o_1) \qquad 1 \le i \le N \\ \psi_1(i) &= 0 \end{aligned} \tag{10}$$

Recursion:

$$\begin{aligned} \delta_t(j) &= \max_{1 \le i \le N}\left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t) \qquad 2 \le t \le T \\ & \qquad\qquad\qquad\qquad\qquad\quad 1 \le j \le N \\ \psi_t(j) &= \text{Argmax}_{1 \le i \le N}\left[\delta_{t-1}(i) a_{ij}\right] \qquad 2 \le t \le T \\ & \qquad\qquad\qquad\qquad\qquad\quad 1 \le j \le N \end{aligned} \tag{11}$$

Termination:

$$p = \underset{1\le i\le N}{\text{Max}}\left[\delta_T(i)\right]$$
$$q_T = \underset{1\le i\le N}{\text{Argmax}}\left[\delta_T(i)\right] \quad (12)$$

Path (state sequence) backtracking:

$$q_t = \psi_{t+1}(q_{t+1}) \quad t = T-1, T-2, \ldots\ldots\ldots, 1 \quad (13)$$

HMMS training algorithm: The training procedure is a variant of a well known K-means iterative procedure for clustering data[2]. We assume we have a training set of observations and an initial estimate of all model parameters. However, unlike the one required for reestimation, the initial estimate can be chosen randomly, or on the basis of any available model which is appropriate to the data. Following model initialization, the set of training observations sequences, is segmented into states, based on the current model λ. This segmentation is achieved by finding the optimum state sequence via the Viterbi algorithm and then backtracking along the optimal path. The resulting of segmenting each of the training sequences is, for each of the N states, a maximum likelihood estimate of the set of the observations that occur within each state S_i according to the current model. An updated estimate of the $b_j(k)$ parameters is:

$b_j(k)$ = number of vectors with codebook index k in the state j divided by the number of vectors in state j.

Based on this segmentation, updated estimates of the a_{ij} coefficients can be obtained by counting the number of transitions from state i to j and dividing it by the number of transitions from state i to any state (including itself). An updated model $\hat{\lambda}$ is obtained from the new model parameters and the Baum-Welch Eq.[13] (for more details) are used to estimate all model parameters. The resulting model is then compared to the previous model (by computing a distance score that reflects the statistical similarity of the HMMs). If the model distance score exceeds a threshold, the old model λ is replaced by the new model $\bar{\lambda}$ and the overall training loop is repeated. If the model distance score falls below the threshold, then the model convergence is assumed and the final model parameters are saved.

Vector quantization: The process of the Vector Quantization (VQ) procedure basically partitions the entire training vectors, equals to 6364 vector in our training set, into M disjoint sets, represents each set by a single vector (V_m, $1\le m\le M$) which is generally the centroid of the vectors in the training set and then iteratively optimizes the partition and the codebook (i.e., the centroids of each partition). Several algorithms exist for designing of an appropriate codebook for quantization, mainly LBG and k-means algorithms, in this study, we used the latter one as we have done in past works[14] and for its universal use[15]. Associated with VQ is a distorsion penalty since we are representing an entire region of the vector space by a single vector. Clearly it is advantageous to keep the distorsion penalty as small as possible. However this implies a large size codebook and that leads to problems in implementing HMMs with a large number of parameters. Once the codebook has been designed, quantization of the input analysis vectors involves computing a Euclidean distance between the input vector and each of the M codebook vectors and assigning the index of the codebook which gave the lowest distorsion to the test frame.

The proposed approach: The class of hidden Markov models which uses a vector quantization on the whole vectors given by the feature extraction step, in our case MFCC vectors with thirteen coefficients, is called discrete HMMs, by contrast in continuous HMMs, a set of vectors belonging to the same HMM state is assumed to have a Gaussian distribution or a mixture of Gaussians. In the former method, we generally quantize a unique vector at time t of dimension D.

If we denote the observation sequence of a word v, ($1\le v\le V$ and V is the set of vocabulary words or digits) by:

$$O_1^T = o_1 o_2 \ldots . o_t \ldots . o_T \quad (14)$$

Where o_t is the observation vector at time t (here a MFCC vector with dimension D).

In the study we proposed, rather than applying the VQ process on vectors of dimension D, we used a concatenation of vectors, which we will call Grouped Vectors (GV) of different lengths, we have used in our study a concatenation of 3,5,7 and 9 vectors increasing the new dimension of the new vectors, respectively to 39,65,91 and 117. By this method, the natural correlation existing between successive frames is implicitly modeled and all the recognition process so far is not changed, since all the steps in the speech recognition system, will perform in the new observation sequence $O^T{}_1'$.
With:

$$O_1^{T'} = o_{1,L}{}' o_{2,L}{}' \ldots . o_{t,L}{}' \ldots o_{T,L}{}' \quad (15)$$

L: length (number) of vectors in the concatenated vector (=3,5,7 or 9).
$O_{t,L}'$: The Grouped Vector (GV) of length L at time t.

$$o_{t,L}{}' = o_{t-[L/2]} \ldots . o_t \ldots o_{t+[L/2]} \quad (16)$$

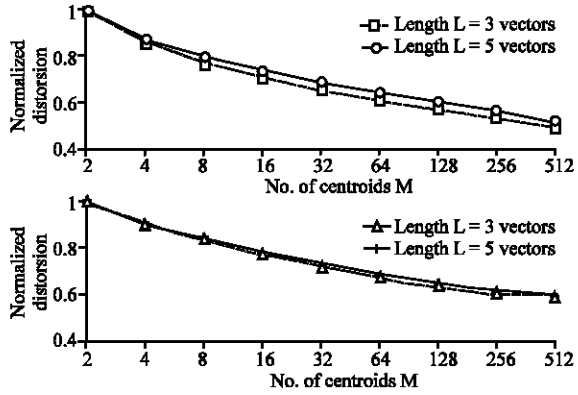

Fig. 2: Curve representing tradeoff of VQ normalized distorsion applied on grouped vectors of length L as a function of the size of VQ codebook

[.] : the floor operator that returns the biggest integer less than its argument.

Because M is not known before running V, an experiment was done to optimally choose it. Figure 2 illustrates the trade-off of normalized quantization distorsion versus different values of M in Grouped Vectors (GV) quantization for L=3,5,7and 9. It can be seen that only decreases in distorsion accrue beyond a value of M=64 for all L (distorsions are between 0.65 and 0.5). We also can notice for a fixed M, distorsions appear very close for L = 3 vs. 5 vectors and for L = 7 vs. 9 vectors. Reduction ratios in distorsion between M = 64 and M = 512 are 18, 20, 13.8 and 11.8%, respectively for L = 3,5,7 and 9. Thus, we considered the size of the VQ centroid M = 64 as a good compromise between the largest of the VQ codebook, the representing ratio (99 rather than 12) and the distorsion penalty reached.

RESULTS AND DISCUSSION

To evaluate our approach described above, a moderate size database was recorded. The task we are dealing with is of recognizing isolated Arabic digits (0-9) in a multispeaker and speaker independent manners. The database was divided into four sets, a Training Set (TnS) consisting of twenty occurrences of each digit by 20 talkers (i.e., a single occurrence of each digit per talker) was used. Half of the talkers were male, half female. For testing, we used three other independent Test Sets (TS) with the following characteristics:

TS-1: The same 20 talkers as were used in the training, 300 occurrences of digits (0-9).

Table 1: Recognition performance (W.E.R) in (%) for the baseline system (HMM/VQ) and the proposed approach (HMM/GVQ)

	HMM/VQ	HMM/ GVQ L=3	HMM/ GVQ L=5	HMM/ GVQ L=7	HMM/ GVQ L=9
TS-1	07.66	08.33	08.33	04.33	04.66
TS-2	13.00	08.66	10.33	11.00	11.00
TS-3	26.76	31.68	28.05	26.10	21.81
Av.	15.80	16.22	15.57	13.81	12.49

TS-2: A new set of 6 talkers, five occurrences per digit per talker were used, giving 300 occurrences of digits.

TS-3: Another new set of 19 talkers, 9 talkers were male, 7 were female, giving 770 occurrences of digits.

In order to see the performance of our proposed approach, called the HMM/GVQ system for HMM Grouped Vector Quantization, that means vectors of different length L grouped in the VQ process, we compare it with the baseline approach called HMM/VQ, in which Vector Quantization is performed only on a single vectors as provided by the feature analysis Fig. 1.

Table 1 presents the results from a series of recognition experiments to determine the effect of adding adjacent vectors to the grouped vector (HMM/GVQ systems). In all the experiments, we trained a single hidden Markov model per digit (0-9), based on a discrete density model, with state observation densities having 64 symbols. Each model was a left-to-right design with 5 states. It can be seen from the table, that the best word error rate WER was 12.49% for the system of the approach proposed where we used 9 vectors to form the new grouped vector GV HMM/GVQ/L=9 and treated as an observation vector, compared to the baseline method HMM/VQ, reduction was about 21%, from 15.8% (HMM/VQ) to 12.49% (HMM/GVQ/L=9). We also can see from the table, that as the grouped vector GV was increased to best model the dependence between vectors, the WER was reduced from 15.57% in the HMM/GVQ system with L = 5 to 12.49% in the HMM/GVQ system with L = 9, except in the case where L equals 3, causing reduction in WER compared to the baseline method, about 1.5, 12.6 and 21%, respectively for HMM/GVQ/L =5, HMM/GVQ/L = 7 and HMM/GVQ/L = 9 systems.

Figure 3 shows a comparison of the results in WER evaluation obtained from the standard method and the proposed one, in multispeaker and speaker independent modes. The baseline system is represented by syst. and syst.1 to syst.4 are for HMM/GVQ/L=3 to HMM/GVQ/L = 9 systems respectively. As depicted in the Figure, in multispeaker mode the best result was given by the proposed approach HMM/GVQ/L = 7 introducing a reduction from 7.66 to 4.33% giving a rate of 43.5 %. In speaker independent mode, where testing sets were

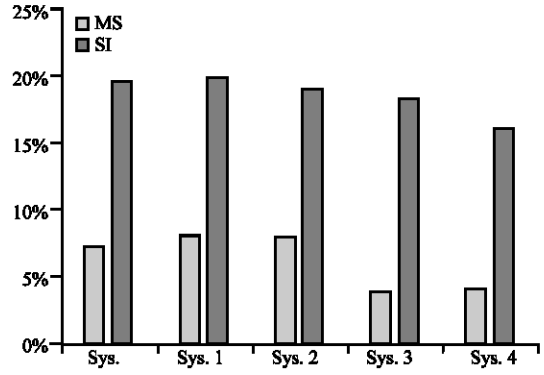

Fig. 3: Comparison performance WER in (%) for the Multi Speaker (SM) and Speaker Independent (SI) modes

constructed from speakers not used in the training set, again the proposed approach brings the best performance, the WER was 16.4% in the HMM/GVQ/L =9 system, rather than 19.88% given by the standard approach, causing a reduction in WER of about 17.5%.

CONCLUSION

The work described in this study is related to a specific task in speech recognition, that is of isolated Arabic digits (0-9) recognition. The use of HMMs for acoustic modeling dominates the field of speech recognition. Although HMMs will continue to play a role in most speech recognition systems for a long time to come, many alternative ideas have been presented in recent years to address some of the shortcomings of HMMs. Thus, we proposed in the study between hand a new approach in dealing with vector quantization and hidden Markov models for speech recognition. We have demonstrated that our method could achieve a significant reduction in word error rate, which is the ultimate goal pursued in speech recognition systems.

ACKNOWLEDGMENT

The authors wish to thank the various members of the automatic and signals laboratory of Annaba for their continuous support and help. I am also grateful to D. Habiba and S. Cherifa for gathering a part of testing set TS-3.

REFERENCES

1. Bahl, L.R., F. Jelinek and R.L. Mercer, 1983. A maximum likelihood approach to continuous speech recognition, IEEE Trans. Patt. Anal. Machine Intl., pp: 179-190.
2. Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition, Proc. IEEE, pp: 257-286.
3. Djemili, R., 2006. Reconnaissance de la Parole Arabe par Modèles Hybrides Markovien Neuro-Prédicteur et Multiclassifieurs Neuronaux, Doctorat Thesis, Badji Mokhtar University, Annaba, Algeria.
4. Lippmann, R.P., 1997. Speech recognition by humans and machines, Speech Communications, pp: 1-15.
5. Tebeleskis, J., 1995. Speech Recognition using Neural Networks, PhD. Thesis, Carnegie Mellon University.
6. Furui, S., 1986. Speaker independent isolated word recognizer using dynamic features of speech spectrum, IEEE Trans. Acoust. Speech Signal Process., pp: 52-59.
7. Wellekens, C.J., 1387. Explicit time correlation in hidden markov models for speech recognition, in Proc. Intl. Conf. Acoust. Speech Signal Process., pp: 384-386.
8. Russel, M., 1993. A Segmental HMM for Speech Pattern Matching, in Proc. Intl. Conf. Acoust. Speech Signal Process., pp: 499-502.
9. Gales, M. and S. Young, 1993. Segmental HMM's for Speech Recognition, in Proc. Euro. Conf. Speech Comm. Tech., pp: 1579-1582.
10. Ostendorf, M., V. Digalakis and O. Kimball, 1996. From HMM's to segment models: A unified view of stochastic modeling for speech recognition, IEEE Trans. on Speech Audio Process., pp: 360-378.
11. Davis, S.B. and P. Mermelstein, 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, in Proc. Intl. Conf. Acoust. Speech Signal Process., pp. 357-366.
12. Forney, G.D., 1973. The Viterbi algorithm, Proc. IEEE, pp: 268-278.
13. Bengio, Y., 1999. Markovian models for sequential data, Neural Computing Surveys, 2: 129-162.
14. Djemili, R., M. Bedda and H. Bourouba, 2004. Arabic Connected Digits Recognition Using the Probabilistic Approach of Hidden Markov Models, in Proc. Intl. Arab Conf. On Inf. Techn., pp: 457-463.
15. Paliwal, K.P. and V. Ramasubramanian, 2000. Comments on modified k-means algorithm for vector quantizer design, IEEE Trans. on Image Process., pp: 1964-1967.